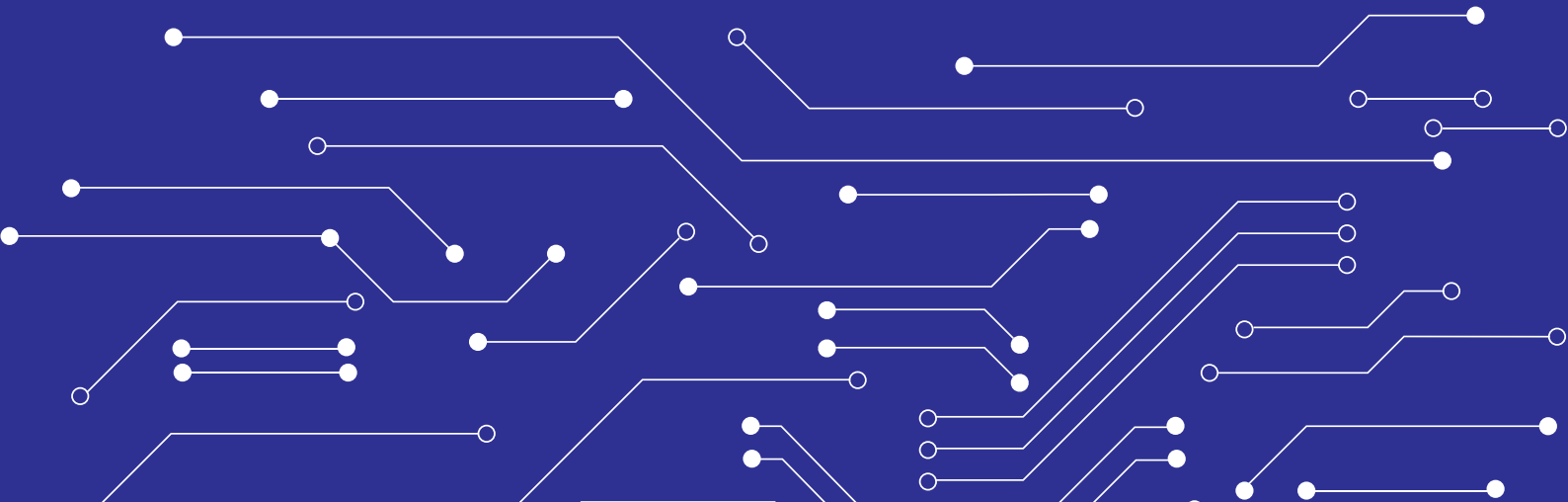
Renovaro cube

Flamingo

Accurate pan-cancer detection model from ultra-low pass whole genome sequencing of cell-free DNA



Executive Summary

Among different modalities for the early detection of cancer, cell-free DNA (cfDNA) extracted via minimally invasive liquid biopsies has emerged as a promising biomarker. As it provides a molecular snapshot of the patient, it is being investigated as a pan-cancer detection biomarker (1).

However, current methods still lack sufficient sensitivity for early detection of cancer (2–4). RenovaroCube ("The Cube") is an AI platform designed to accelerate cancer diagnostics. It is our firm belief that no single model or modality will reach the requisite sensitivity to detect cancer early.

Therefore, The Cube integrates multi-omic data, offering a uniquely comprehensive approach to cancer detection by leveraging a library of trained models for multiple omic layers. Here, one such underlying model is presented, focusing on the detection of cancer from cfDNA sequencing data using fragmentomics.

The mapping coordinates from as few as 200,000 cfDNA fragments per patient were extracted from the publicly available Delfi cfDNA WGS experiments (5,6).

Fragment lengths and sequence motifs of each fragment endpoint were inferred and the Shannon entropy of the sequence motifs was calculated.

The normalised motif and length counts, along with the Shannon entropy, were used as input for Flamingo: a carefully designed neural network trained to differentiate cancer from healthy samples. Performance assessment was conducted through five iterations of 10-fold-cross-validation on the training cohort, supplemented by an evaluation on a withheld validation cohort. Utilising the Delfi dataset comprising healthy donors (n=136) and cancer patients (n=127) spanning 7 different cancer types (bile duct cancer (n=10), breast cancer (n=23), colorectal cancer (n=15), gastric cancer (n=9), lung cancer (n=40), ovarian cancer (n=14), and pancreatic cancer (n=16)), Flamingo demonstrated robust performance, achieving a median area under receiver-operator-curve (AUROC) of 0.952. At a predetermined specificity of 98%, Flamingo demonstrated an overall sensitivity of median 74.2% for detecting cancer during the cross-validation.

In the separate, unseen validation set, containing 55 healthy donors and 48 cancer patients, Flamingo maintained its performance, achieving 98.2% specificity and 75.0% sensitivity with an AUROC of 0.966.

This confirms the model's capacity to generalise beyond the training dataset, demonstrating its potential to be robust in diverse clinical scenarios.

Moreover, Flamingo and the original Delfi model each identified non-overlapping cancer cases not detected by the other, suggesting that some cancer patients manifest distinct cancer-related signals.

Consequently, it is apparent that no single model will suffice for detecting all cancers. Instead, the collective insights from multiple models and modalities may provide the necessary discriminatory power for application in a real-world screening setting.

In conclusion, utilising only minute amounts of cfDNA WGS data, we were able to develop Flamingo, a model that performs on par with state-of-the-art cfDNA cancer detection models.

Integrating Flamingo into The Cube platform reinforces its pool of models operating across diverse omic layers, to detect different cancer types as early as possible.

The Cube provides clinicians with the tools needed to identify malignancies at their earliest occurrence, thus facilitating timely therapeutic interventions that can lead to improved patient outcomes.

Flamingo Development

As few as only 200,000 fragments per patient were extracted from the Delfi dataset (5), available on the public database FinaleDB (6). For each fragment, we extracted the length, and the sequence motifs near the 5' endpoints.

A weighting factor was applied to each motif, corresponding to the probability of that fragment's length being derived from a tumour cell vs a healthy cell (7).

The motifs and fragment lengths were used as input for separate arms of a neural network, named FLAMINGO: Fragment Lengths And Motifs In Neural Grouped Observations.

Flamingo's performance was first evaluated in five times repeated 10-fold-cross-validation.

For each repeat, we recorded the area under receiver-operator-curve (AUROC), and the sensitivity at both 98% and 99% specificity.

For clinical settings, such as early cancer detection in a screening population, a high specificity is essential to minimize the number of individuals unnecessarily referred for follow-up diagnosis.

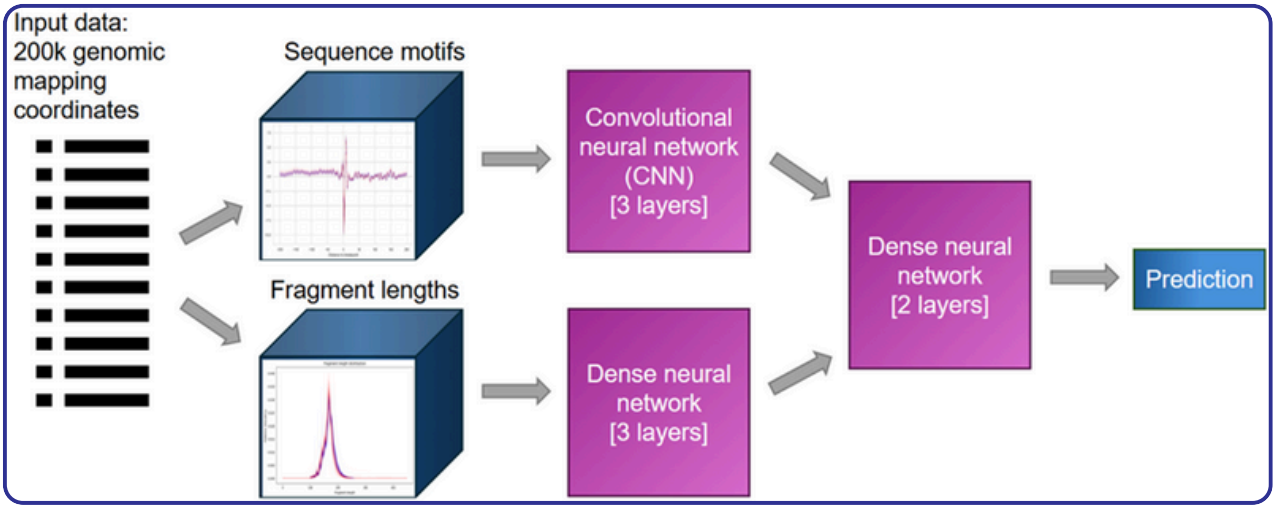


Figure 1. *Schematic outline of Flamingo workflow and data streams.*

Flamingo Perfomance

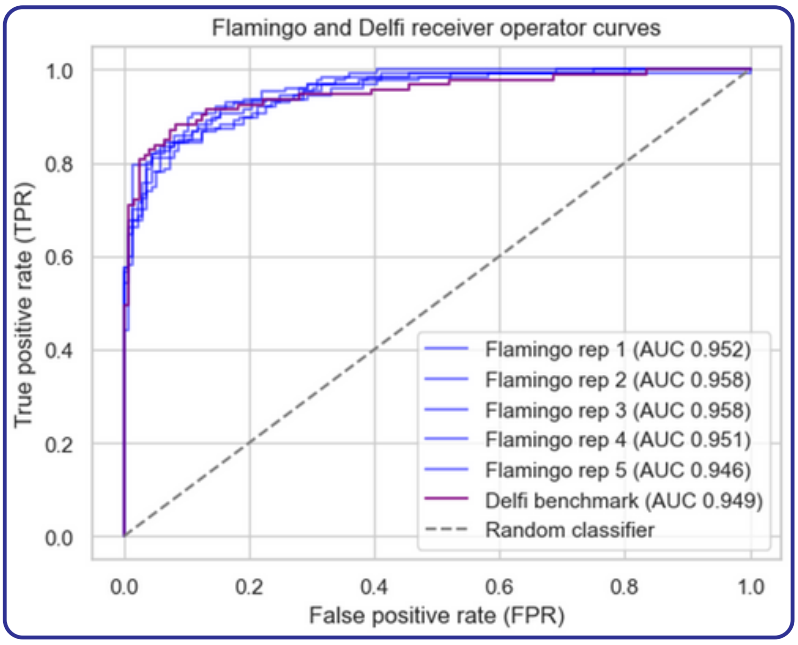


Figure 2. *ROC curves for the five repetitions of 10-fold-cross-validation. TPR: true positive rate; FPR: false positive rate.*

Across repetitions, the median AUROC was 0.952 (95% CI: 0.948 - 0.959) and the median sensitivity was 74.2% (95% CI: 69.5 - 80.1%) and 73.1% (95% CI: 68.3 - 74.7%) at 98 and 99% specificity, respectively.

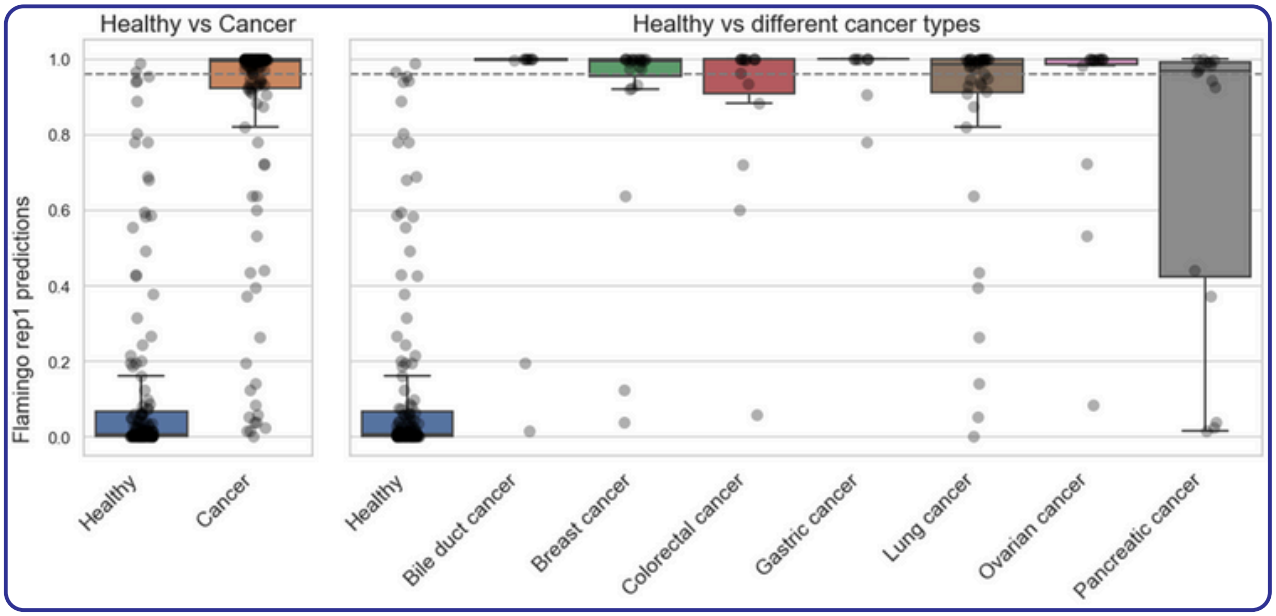


Figure 3. *Flamingo prediction scores from the repeat with median performance (repeat #1 in figure 2), represented in boxplots. Left: Comparison of the distribution of the Flamingo prediction scores between cancer patients and healthy donors. Right: Comparison of the Flamingo prediction scores across different cancer types and healthy donors. The grey dashed lines indicate the 99% specificity cutoff.*

Comparison w/ Delfi

Delfi is a state-of-the-art model for the detection of cancer signals from cfDNA using fragmentomics (5). Briefly, it finds the ratio between short (100-150 bp) and long (151-250 bp) cfDNA fragments on 50 Mb bins across the genome. Using LASSO regression and principal component analysis (PCA) on 473 features, they were able to distinguish various cancer types from healthy individuals.

A head-to-head comparison between Delfi and Flamingo was possible using the published Delfi prediction scores per sample. The prediction scores were available for only a subset of the data, limiting the analysis to 121 healthy samples and 93 samples from patients with cancer.

In this comparison, it is worth noting that Delfi uses on average 46.8 million sequenced fragments per sample (ranging from 1.3 to 224.8 million). In contrast, Flamingo only uses 0.2 million fragments, or 0.4% of the data, which reduces sequencing costs, data storage footprint, and analysis time.

Despite the stark difference in data usage per sample, performance is comparable between the two models. Flamingo's median AUC of 0.952 matches Delfi's AUC of 0.949.

Similarly, the sensitivity at high specificity cutoffs of 98% (sensitivity 74.2%) and even 99% (sensitivity 73.1%) match those of Delfi at 76.3% and 71.0%, respectively.

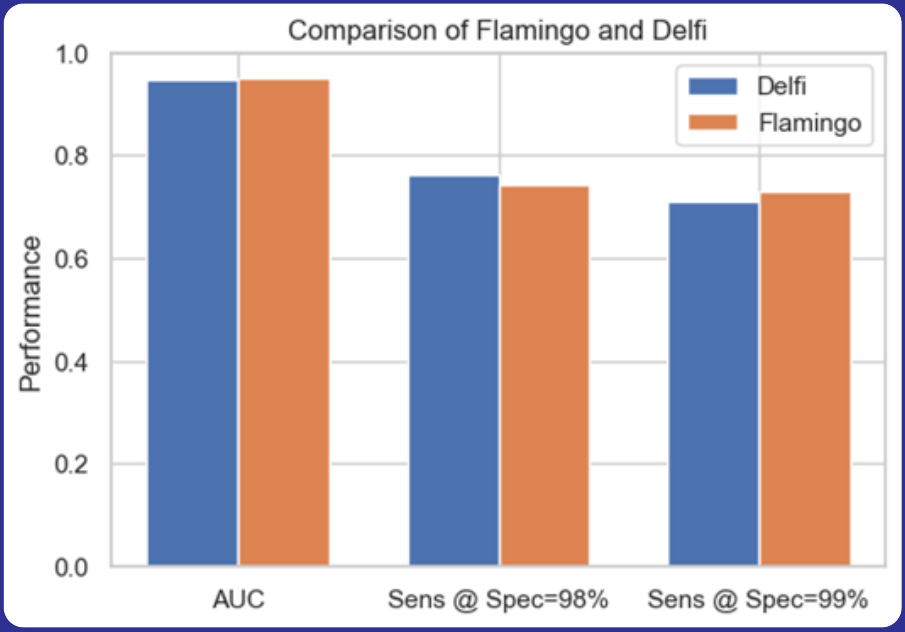


Figure 4. Performance comparison between the Delfi (blue) and Flamingo (orange) models. The results are highly comparable for AUC, and sensitivity at 98% and 99% specificity.

Comparison w/ Delfi

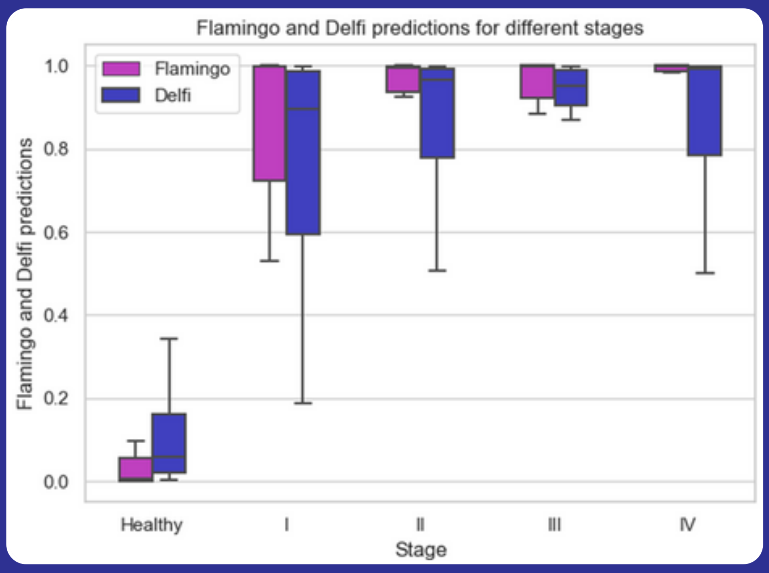


Figure 5. Prediction scores for Flamingo and Delfi across cancer stages.

Predictions at various stages of cancer are similarly comparable between Delfi and Flamingo, as depicted in figure 5.

Importantly, the two models detected partially non-overlapping cancer cases, as illustrated in figure 6. Cancer samples in the upper left quadrant (n=12) were detected by Delfi and missed by Flamingo, while cancer samples in the lower right quadrant (n=14) were detected by Flamingo and missed by Delfi. No correlation was found for cancer type or cancer stage for these non-overlapping cases, indicating that any cancer type or stage is equally likely to be missed by one model or the other.

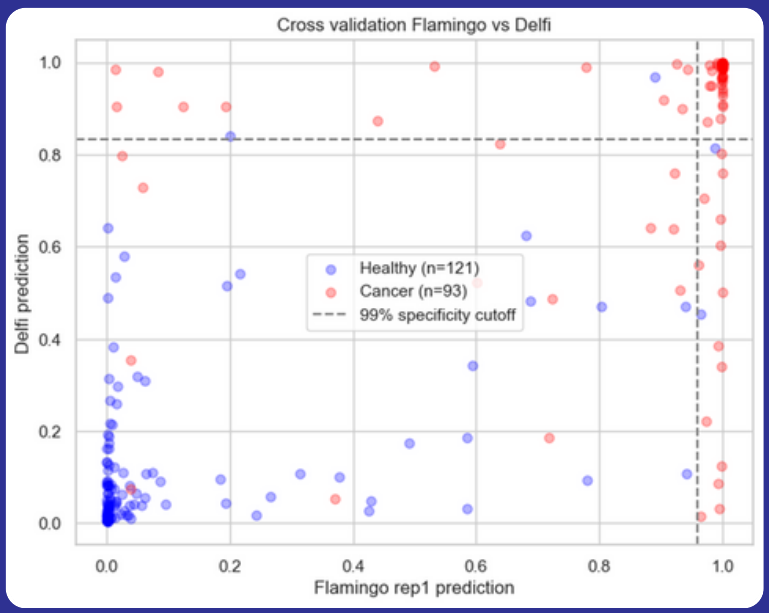


Figure 6. Comparison between Flamingo and Delfi predictions. Cancer prediction scores from the Flamingo model (x-axis, results from median repeat depicted) and the Delfi model (y-axis). Grey dashed lines indicate the 99% specificity cutoffs.

The Cube

The complementarity of Flamingo and Delfi illustrates that different models detect diverse types of cancer signals, even from the same input data. This reinforces the idea that no singular model may be sufficient to detect all distinct types of cancer signals, and that a multi-omic platform including numerous models will be required for robust and sensitive detection of cancer, especially in the setting of early detection.

The Cube is a unique artificial intelligence (AI) platform that integrates multi-omic data to infer the presence of cancer.

On the one hand it powerfully leverages decades worth of tissue-based research to identify cancer-related signatures in e.g. methylation, copy number variations (CNV), and gene expression to find consistent patterns in cfDNA.

At the same time, The Cube includes liquid-only fragmentomics-based models like Flamingo.

The more data, and the more models that are fed into The Cube, the better its predictions become.

Discussion & Conclusions

There is an unmet need for robust and sensitive early detection of cancer, and early detection of the recurrence of cancer after treatment.

Liquid biopsies in the form of cfDNA extracted from blood plasma hold the potential to meet this need.

However, current methods based on singular omic layers and stand-alone models still lack adequate sensitivity for application in routine screening settings.

Here, we propose The Cube as an AI platform that integrates numerous models across multiple omic layers. As a case in point, we developed Flamingo, a cfDNA fragmentomics-based model that incorporates fragment lengths and sequence motifs. Flamingo performs on par with state-of-the-art fragmentomics models, using only a fraction of the data that the competitor uses. Flamingo will be an asset to The Cube, adding to the pool of models employed in the detection of cancer.

Authors

Daan Vessies

Molecular Biology

Daan studied Life Science & Technology at TU Delft and Leiden University. Following his internship and several years of work at the Genomics Core Facility of the Netherlands Cancer Institute (NKI), he joined dr. Daan van den Broek's lab (NKI) to pursue his PhD. His research focuses on enhancing circulating tumour DNA (ctDNA) detection methods. Daan's primary interest lies in fragmentomics, where he aims to develop and refine methods for the highly sensitive detection of cancer in challenging clinical settings.

Edward Post

Molecular Biology

Started as a research technician for the molecular pathology at the Erasmus MC, designing sequencing assays to improve cancer diagnostics. He then joined thromboDx, a startup focused on a pancancer detection assay using platelet derived RNA. After 8 years of being a technician and slowly growing into the role of bioinformatician. Completing his PhD in Bioinformatics at the Amsterdam UMC. He specializes in liquid biopsies and bioinformatics, with most of my experience in the analysis of NGS data

Teresa Bucho

Biomedical Engineer/Deep Learning

Mrs Bucho is a Biomedical Engineer. After finishing her masters in Lisbon, she moved to Amsterdam to pursue a PhD at the department of Radiology at the Netherlands Cancer Institute. During her PhD, she analysed the limitations of current response assessment criteria to treatment in solid tumors (RECIST) and explored how ML/DL could be used to overcome some of these limitations. Specifically, she worked on a model to estimate the entire tumor burden of a patient from CT scans and on a model to track cancer lesions longitudinally between baseline and follow-up scans. Currently she is working on exploring how imaging and deep learning can be incorporated into the projects and its added value.

Authors

Dennis Makarawung

Medical Director

Dennis holds a medical degree from the University of Amsterdam. During his academic years, he successfully completed a research internship at Harvard Medical School. Subsequently, Dennis embarked on a career as a surgery resident in various academic and teaching hospitals, and several private clinics in the Netherlands while concurrently undertaking a PhD project centred around obesity and weight loss interventions.
Throughout his career, Dennis developed a keen interest in prevention and early detection, prompted by the often symptomatic approach observed in current healthcare practices.

Frank van Asch

Chief Technology Officer

Frank holds a Postgraduate Degree in Business Analytics and Data Science. He is multi-disciplinary skilled in Data Science and Data Mining with a track record in FinTech, HighTech and BioTech. Domain knowledge and experience of existing and emerging technology and platforms especially but not limited to data science, advanced analytics and big data. As CTO, he is responsible for the architecture and design of the GEDi platform, proprietary algorithms and validated machine learning models.

Bibliography

1. Campos-Carrillo A, Weitzel JN, Sahoo P, Rockne R, Mokhnatkin J V., Murtaza M, et al. Circulating tumor DNA as an early cancer detection tool. Pharmacol Ther. 2020 Mar;207:107458.

2. Fernandez-Uriarte A, Pons-Belda OD, Diamandis EP. Cancer Screening Companies Are Rapidly Proliferating: Are They Ready for Business? Cancer Epidemiology, Biomarkers & Prevention. 2022 Jun 1;31(6):1146–50.

3. Duffy MJ, Crown J. Circulating tumor DNA (ctDNA): can it be used as a pan-cancer early detection test? Crit Rev Clin Lab Sci. 2023 Nov 7;1–13.

4. Pons-Belda OD, Fernandez-Uriarte A, Diamandis EP. Can Circulating Tumor DNA Support a Successful Screening Test for Early Cancer Detection? The Grail Paradigm. Diagnostics. 2021 Nov 23;11(12):2171.

5. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019 Jun 29;570(7761):385–9.

6. Zheng H, Zhu MS, Liu Y. FinaleDB: a browser and database of cell-free DNA fragmentation patterns. Bioinformatics. 2021 Aug 25;37(16):2502–3.

7. Vessies DCL, Schuurbiers MMF, van der Noort V, Schouten I, Linders TC, Lanfermeijer M, et al. Combining variant detection and fragment length analysis improves detection of minimal residual disease in postsurgery circulating tumour <scp>DNA</scp> of stage <scp>II–IIIA NSCLC</scp> patients. Mol Oncol. 2022 Jul 27;16(14):2719–32.

Renovaro cube

RenovaroCube is a leading molecular data science company pioneering cancer detection and treatment through its advanced AI/Machine Learning platform. Leveraging multi-omics analysis and proprietary algorithms, we developed over 3600 biomarker panels for precision oncology, supporting early diagnosis and personalized treatment planning. Our platform enables predictive responses to therapies, minimal residual disease monitoring, and recurrence prevention.

Committed to early cancer detection, we aim to increase treatment success rates by identifying cancer before symptoms appear. Our journey from FinTech to oncology innovation has led to groundbreaking discoveries and strategic partnerships with academic institutions, healthcare organizations, and industry leaders.

Driven by a multidisciplinary team of experts, including molecular biologists, biomedical scientists, doctors, software engineers, and data scientists, we're dedicated to pushing the boundaries of cancer research. RenovaroCube empowers biomedicine with big data science, offering hope to patients and clinicians through transformative technologies.

Fred. Roeskestraat 115,
1076 EE Amsterdam
karen@renovarocube.com
renovarocube.com